DOCUMENT RESUME

ED 275 751 TM 860 668

AUTHOR Blair, R. Clifford; Higgins, James J.
TITLE On the Relative Power of the Paired Samples t Test and Wilcoxon's Signed-Ranks Test.
PUB DATE Apr 85
NOTE 75p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, Chicago, IL, March 31-April 4, 1985).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Estimation (Mathematics); Literature Reviews; Measurement Techniques; *Monte Carlo Methods; *Nonparametric Statistics; *Sample Size; *Statistical Distributions; Statistical Studies; *Test Theory
IDENTIFIERS Chi Square; Power (Statistics); *T Test; *Wilcoxon Matched Pairs Signed Ranks Test

ABSTRACT

Monte Carlo methods were employed to assess the relative power of the paired samples t test and Wilcoxon's signed-ranks test under ten population shapes. Results of the study indicated that: (1) each of the two statistics was more powerful than the other in given situations; (2) the power advantages of the t test under normal theory were small; (3) in the nonnormal situation the t test never attained more than modest advantages over the Wilcoxon test; (4) the Wilcoxon test was not only more often the more powerful test, but in those situations where it was more powerful, it was often vastly so; and (5) the magnitude of the Wilcoxon's power advantage often increased with increases in sample size. It was concluded that, insofar as these two statistics are concerned, the often repeated claim that parametric tests are more powerful than nonparametric tests is not justified. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED275751

On The Relative Power Of The Paired Samples *t* Test And Wilcoxon's Signed-Ranks Test

R. Clifford Blair

Department of Measurement and Research

University of South Florida

James J. Higgins

Statistics Department

Kansas State University

Running head: ON THE RELATIVE POWER

Paper presented at the 1985 annual meeting of AERA, Chicago, March 31-April 4.

BEST COPY AVAILABLE

TM 860 668

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R.C. Blair

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Abstract

Monte Carlo methods were employed to assess the relative power of the paired samples *t* test and Wilcoxon's signed-ranks test under ten population shapes. Results of the study indicated that: (1) each of the two statistics was more powerful than the other in given situations, (2) the power advantages of the *t* test under normal theory were small, (3) in the nonnormal situation the *t* test never attained more than modest advantages over the Wilcoxon test, (4) the Wilcoxon test was not only more often the more powerful test, but in those situations where it was more powerful, it was often vastly so, and (5) the magnitude of the Wilcoxon's power advantage often increased with increases in sample size. It was concluded that, insofar as these two statistics are concerned, the often repeated claim that parametric tests are more powerful than nonparametric tests is not justified.

On The Relative Power Of The Paired Samples *t* Test And Wilcoxon's Signed-Ranks Test

Authors of statistics and research methodology texts designed for use by students and research workers in the social and behavioral sciences often discourage the use of nonparametric tests as general analytic procedures. Instead, these authors usually relegate the use of these techniques to the somewhat infrequently (in their opinion) encountered situation in which the robustness of the usual parametric procedure cannot be relied upon to counteract the effects of one or more serious violations of its underlying assumptions. This sort of situation might occur, for example, when sample sizes are quite small and data are sampled from markedly nonnormal (e.g. highly skewed) populations.

This being the case, readers are often admonished to employ parametric techniques whenever possible. The following statements are typical of such recommendations. "Where there is any choice...we should prefer a parametric test, except where a quick, rough test will do" (Guilford and Fruchter, 1978; p. 212). "Unless there is good evidence to believe that populations are rather seriously non-normal and that variances are heterogeneous, it is usually unwise to use a nonparametric statistical test in place of a parametric one" (Kerlinger, 1964; p. 259).

The primary source for the above recommendations stems from certain commonly held beliefs regarding the relative power/efficiency of parametric and nonparametric tests. The following

statements taken from an available sample of textbooks will provide some insights into these beliefs. Guilford and Fruchter (1978) preface their admonition concerning the use of nonparametric tests with the following. "Before an investigator resorts to them (i.e. nonparametric tests), however, he should consider whether any of the more powerful [i.e. parametric tests] can be used" (italics added). Kerlinger (1964) justifies his recommendation quoted above with the following statement. "The reason for this is that parametric tests are almost always more powerful than nonparametric tests" (italics added).

Other authors are of the same mind. For example Kurtz (1983, p. 160) states, "...nonparametric tests should be used only when necessary, because they are somewhat less sensitive to differences than their parametric alternatives." Johnson (1977, p. 190) agrees with this point of view stating, "Sometimes it appears that data can be analyzed by both types of tests. When this occurs, preference should be given to the parametric test. It has greater power and is more likely to detect departures from the null hypothesis." Couch (1982, p. 327) writes, "In most cases the parametric statistical tests have more statistical power than their nonparametric counterparts." The same author concludes, "This makes the nonparametric tests a good choice for a 'quick look' at the data." A look in the subject index of the popular text by Hopkins and Glass (1978) reveals but a single reference to nonparametric statistics. This reference appears in the midst of a footnote where the authors state simply,

"...nonparametric statistics-methods which make fewer assumptions but are also less efficient." Other authors omit any mention of nonparametric or distribution-free tests.

Statements similar to those quoted above are found quite commonly in social and behavioral science statistics and research methodology textbooks. (See for example: Best (1981, p. 287), Chase (1976, p. 211), Cornell (1956, p. 238), Downie & Starry (1977, p. 241), Engelhart (1972, p. 278), Garrett (1966, p. 267), Gay (1976, p. 251), Klugh (1970, p. 306), Kurtz & Mayo (1979, p. 393), Mendenhall (1971, p. 369), Minium (1970, p. 315), Nunnally (1975, p. 291), Pagano (1981, p. 398), Roscoe (1975, p. 22), and Spence, Underwood, Duncan & Cotton (1967, p. 209).)

Several interesting points should be noted in regard to the above citations. (1) All the authors assert that nonparametric tests are less powerful (or less efficient) than are parametric tests. (2) These assertions are either unqualified or are only vaguely qualified. (3) No references are given to support these statements thereby leading the reader to believe that this information falls into the realm of "common knowledge". (4) None of the authors cited raise the possibility that nonparametric tests may be as powerful as, or more powerful than, their parametric counterparts.

In contrast to the above, some authors do qualify their assertions of parametric power superiority by pointing out that this result is necessarily realized only when the assumptions underlying the parametric tests are (perfectly) met. Examples of

such statements can be found in Ary and Jacobs (1976, p. 420), Elifson, Runyon and Haber (1982, p. 423), Hays (1963, p. 618), Hinkle, Wiersman and Jurs (1982, p. 333), Kirk (1968, p. 503) and Sharp (1979, p. 145). With the exception of Elifson, Runyon and Haber, these authors, like the first group, fail to mention the possibility that nonparametric tests may be as powerful as parametric tests. None of these authors raise the possibility that nonparametric tests may be more powerful than their parametric counterparts. Like the first group, these authors do not reveal the sources from which their statements are derived.

Of the twenty-nine textbooks reviewed for this paper (actually, more were reviewed but failed to mention nonparametric or distribution-free tests), only one raised the possibility that nonparametric tests may attain efficiencies superior to those of their parametric counterparts. In this regard, Downie and Heath (1970, p. 265) state, "In terms of a mathematical criterion of statistical efficiency, distribution-free tests are often superior or equal to their parametric counterparts when the assumptions of the nonparametric test are met, but the assumptions of the parametric test are not." These authors cite Bradley (1968) as the source for this statement. Bradley in turn cites sources to be discussed below.

The Present Study

Purpose

With assertions of parametric power superiority being so plentiful, one might assume that the evidence for such statements

is both abundant and convincing. This is not, generally speaking, the case. This latter statement remains true even when attention is restricted to the more familiar procedures.

The purpose of this study is to supplement the (as will be demonstrated) inadequate knowledge base concerning the relative power of two of the more familiar statistical tests; the parametric paired samples t test, and the nonparametric Wilcoxon signed-ranks (WSR) test. The relative power of these two statistics will be examined by means of a large scale Monte Carlo study. Before describing the methods and results of this study, however, we shall review the existing literature pertaining to the relative power/efficiency of the two statistics in question. This review should prove interesting since virtually every author cited above included these two statistics in their presentations of nonparametric tests.

A Review of the Literature Pertaining to the Relative Power/Efficiency of the Paired Samples t Test and Wilcoxon's Signed-Ranks Test

With these two statistics being so well known and since they are often presented as "counterparts" or even "rivals", one might assume that their power/efficiency properties have been carefully and extensively compared. This is not the case. In fact, the researcher who turns to the literature for guidance in choosing the more powerful/efficient of these two statistics in non-ideal (i.e. applied) situations is almost certain to come away disappointed and/or confused. Before turning to this literature, it

will be helpful to briefly review the methods that have been employed to compare the efficiencies of the two statistics of interest.

Perhaps the two most common techniques used by statisticians to compare the efficiencies of two statistical tests are those referred to as "relative efficiency" (R.E.) and "asymptotic relative efficiency" (A.R.E.). A lesser known technique is termed "Bahadur relative efficiency" (B.R.E.). A.R.E. and B.R.E. are often referred to as "large sample efficiencies" while R.E.s because of the way they are often used, are termed "small sample efficiencies".

Bradley (1968, p. 57) defines relative efficiency as follows. The relative efficiency of test A with respect to test B is generally defined as b/a, where a is the number of observations required by test A to equal the power of test B based on b observations, when both statistics test the same null hypothesis Ho against the same alternative hypothesis Ha at the same significance level α (both either one-tailed or two-tailed)." The same author offers this (nonmathematical) definition for A.R.E.

> Let A and B be two tests based on a and b observations respectively, each test statistic being asymptotically normally distributed...and each testing the same null hypothesis Ho against the same class of one-sided alternatives Ha>Ho, against which both tests are consistent. The A.R.E. of A with respect to B is the limiting value of the ratio b/a as a is allowed to vary in such a way

as to give A the same power as B while simultaneously

b approaches infinity and Ha approaches Ho.

For a more rigorous (i.e. mathematical) definition see Appendix 6 of Lehmann (1975).

If a random sample is taken from a population having a continous, symmetric density function f(x), then the A.R.E. of the W.S.R. test relative to the t test is

$$ARE = 12\sigma^2[\int f^2(x)]^2$$

when testing for shift alternatives (Lehmann, 1975; pp. 175 and 376). Interestingly, this formula also applies to Wilcoxon's rank sum test (which is mathematically equivalent to the Mann-Whitney U test) and the independent samples t test.

Though not nearly so well known as R.E. and A.R.E. by researchers in the social and behavioral sciences, B.R.E. has also been used as a mechanism to compare the efficiencies of the WSR and t tests. The B.R.E. of two test statistics is obtained by comparing the exponential rate of convergence to zero of the Type I error rate while keeping the Type II error rate fixed (Bahadur, 1960; Klotz, 1965). One can think of B.R.E. as a means of comparing the p-values of two tests. A major advantage of this means of efficiency assessment as compared to A.R.E. lies in the fact that B.R.E. can be defined for explicit shift alternatives of various magnitudes. It will be remembered that A.R.E. is computed under conditions in which the shift alternative must approach the null value of zero as sample sizes increase in order to maintain fixed values of α and β. With these brief definitions in mind (the

reader who is unfamiliar with these methods should see the cited references), we turn our attention to the literature dealing with the issues at hand.

Table 1, developed from results given by Elashoff and Elashoff (1978), Lehmann (1975) and Siddiqui and Raghunandanan (1976) shows A.R.E.s calculated for the WSR test relative to the t test under various symmetric distributions. The notation associated with the mixed normal distributions indicates that these distributions are formed in the following manner: (1) An observation is randomly drawn from one of two normal distributions, (2) with probability 1-T that this observation is randomly drawn from a normal distribution with mean μ and variance σ^2, (3) with probability T, that this observation is randomly drawn from another normal distribution having mean μ and variance $k^2\sigma^2$ (Elashoff & Elashoff, 1978, p. 231). The Student's t distributions are simply t distributions with the indicated degrees of freedom.

Insert Table 1 About Here

The entry of .96 in this table is indicative of a small A.R.E. advantage favoring the t test when the distribution is normal. This result is consistent with the uniformly most powerful (UMP) status obtained by the t under normal theory. Bradley (1968, p. 99) has noted that under normal theory assumptions and under the same conditions necessary for the calculation

of A.R.E., as sample size diminishes from infinity toward zero the efficiency of the WSR test relative to the t test increases from .96 and approaches unity. The magnitudes of these results are interesting since the statements of parametric power superiority noted earlier probably derive from normal theory considerations.

The table entry of 1.00 indicates that neither test holds an advantage under the light-tailed uniform (or rectangular) distribution. Some readers may be surprised to note that the nonparametric test not only gains the advantage under the heavier tailed distributions but also that the magnitude of this advantage rapidly increases with increases in tail weight until reaching infinity under the extremely heavy-tailed Cauchy distribution (a t distribution with 1 degree of freedom).

An interesting and potentially important result was obtained by Hodges and Lehmann (1956) who showed that so long as the difference score population is symmetric and continuous, its shape being otherwise unspecified, the A.R.E. of the WSR relative to the t test may be as high as infinity (as we just saw for the Cauchy distribution) but may never fall below a lower bound of .864. Commenting on this result Hodges and Lehmann (1956) state (p. 356):

> To the extent that the above concept of efficiency adequately represents what happens for the sample sizes and alternatives arising in practice, this result shows that the use of the Wilcoxon test instead of the

> Student's t test can never entail a serious loss of efficiency for testing against shift. (On the other hand, it is obvious...that the Wilcoxon test may be infinitely more efficient than the t test.)

Klotz (1965) has computed B.R.E.s for the WSR test relative to the t test under a normal distribution with variance equal to one. Table 2 gives several of these B.R.E.s as calculated for various shift alternatives. For values of μ near zero, B.R.E. and A.R.E. yield very similar results. As μ increases, B.R.E. increases until μ reaches a value of about .75. At this point

---

Insert Table 2 About Here

B.R.E. attains a high of approximately .981. AS μ increases above .75, B.R.E. declines toward a value of zero at $\mu=\infty$. It can be seen therefore, that by this means of comparison, the magnitude of the treatment effect plays an important role in determining the relative efficiency of these two statistics.

Klotz (1963), using theoretical methods, obtained small sample efficiencies of the WSR test relative to the t test for normal shift alternatives. To compute efficiency Klotz found a sample size n^*, possibly a fractional value determined by interpolation, such that the t test would have power equivalent to that of the WSR test for a given n, σ and mean. The efficiency was then defined to be n^*/n. Table 3 gives selected results from this study. The magnitudes of these small sample

---

Insert Table 3 About Here

---

efficiencies are quite compatible with the A.R.E. of .955 that is obtained for these two tests under normal theory assumptions.

Arnold (1965) found that the small sample power of the WSR test deteriorates badly in comparison to the best rank statistic when samples are taken from certain heavy-tailed distributions. However, the WSR test was still more powerful than the t test in these situations. The distributions considered in this study were t distributions with 1/2, 1, 2 and 4 degrees of freedom. Selected results from the Arnold study are given in Table 4.

Insert Table 4 About Here

This table depicts the power of the sign, WSR and t tests for given sample sizes and values of α under the Cauchy distribution. Again, some readers may be surprised to note that it is the sign and t tests that show respectively the most and least power in this situation.

Thompson, Govindarajulu and Doksum (1967) report relative power calculations for the WSR and t tests obtained for a sample size of 20 drawn from normal and double exponential (Laplace) distributions. In the normal situation the relative power ranged from .93 to .99 while in the double exponential case the calculations ranged from 1.01 to 1.15 (Elashoff & Elashoff, 1978). This indicates a small disadvantage for the WSR test under normality

which turns into a small advantage under the double exponential distribution.

Some Comments on the Literature

The literature just reviewed does not support the belief that the parametric paired samples t test is, in any general sense, more powerful or efficient than the nonparametric WSR test. Asymptotic (i.e. A.R.E.) as well as small sample results indicate that the t test attains only minor advantages over the WSR test under the normal distribution. On the other hand, when the normality assumption is relaxed the WSR test can attain truly large (in theory, infinitely large) advantages over the t test. At the same time it must be noted that B.R.E.s indicate that the t test may attain truly impressive advantages over the WSR when, under normal theory, the shift alternative is quite remote from the null distribution. In this situation the B.R.E. of the two statistics approaches zero as the alternative approaches infinity thereby suggesting an infinitely large advantage for the t test.

But what do all of these theoretical results portend for the practitioner? More specifically, can these rather confusing results be relied upon to guide the researcher in choosing between these two statistics in applied situations? The following quotation from Klotz (1963, p. 631) would indicate that at least one statistician has arrived at a firm conclusion.

> Because of the extremely high efficiency of the nonparametric tests relative to the t in the region of

interest, it is the author's opinion that the nonparametric tests would be preferable to the *t* in almost all practical situations. The exactness of the null distribution, good power for a wide class of shift alternatives, and the negligible loss in efficiency on the home ground of the *t* test support this conclusion.

Other statisticians restrict their endorsements to specific situations. For example, Hettmansperger and McKean (1978) state, "It is by now well documented that for heavy tailed populations, the *t*-test and sample mean are much less powerful and efficient than the Wilcoxon test and the median of pairwise averages." These authors cite Lehmann's book (Lehmann, 1975) as authority for their statement. The portions of Lehmann's book that deal with the subject at hand have been largely dealt with here in the review of literature.

Elashoff and Elashoff (1978, p. 236) writing in the International Encyclopedia of Statistics conclude, "The preceding validity and efficiency studies suggest that if the underlying symmetrical distribution might have longer tails than the normal, the Wilcoxon test or trimmed *t* test would be preferable to the Student's *t* test." The articles alluded to by these authors have also been dealt with here in the review of literature.

Other writers have voiced strong reservations concerning the usefulness of the asymptotic methods employed in these studies. These concerns are particularly important in this context since

the bulk of information regarding the relative efficiency of the WSR and t tests is in the form of asymptotic results.

James Bradley has noted several serious difficulties associated with the use and interpretation of A.R.E.s. Among them he cites the following (Bradley, 1968; p. 58).

> But they [A.R.E.s] suffer from the serious disadvantage that two of their qualifying conditions are totally unrealistic to experimental practice. A.R.E.s represent the relative efficiency of the test under the conditions...that sample size is infinite, that the alternative hypothesis is essentially identical to the null hypothesis, and that the test is one-tailed... No experimenter takes infinitely large samples, and virtually no one is interested in power to reject hypotheses that differ only infinitesimally from the null hypothesis. Furthermore, since the power of any consistent test approaches 1.00 as sample size approaches infinity and since A.R.E. applies only to consistent tests in the case of infinite sample sizes, it follows that a test with a very low but nonzero A.R.E. has as much power at infinite n as could be desired to reject <u>any</u> fixed $H_a \neq 0$.

Mendenhall (1971, p. 373) is more direct in his assessment of A.R.E. This statistician writes, "This quantity [A.R.E.] is, indeed, compact, but is meaningless because the experimenter is not concerned with minute departures from the null hypothesis

and actually is not dealing with infinitely large samples." This author goes on to point out that other asymptotic methods suffer from similar difficulties. If the researcher who turns to the literature accepts these criticisms of A.R.E. (and implicitly B.R.E. as well) then she/he must also be prepared to dismiss as irrelevant, or at least label as highly questionable, most of what the literature has to say regarding the relative efficiency of the two tests of interest.

Small sample efficiency tends to be highly particular in terms of sample size, nature of the alternative hypothesis and significance level. In addition, Bradley (1968, p. 57) has pointed out that small sample relative efficiencies are also highly peculiar to the mathematical methods used to derive them. Other perfectly reasonable definitions of relative efficiency, based on slightly different mathematical procedures may lead to quite contrasting results. Thus, the few small sample results that are available do not constitute a very firm basis for drawing conclusions regarding the two tests of interest.

In light of the literature review and the succeeding comments, it seems fair to say that relatively little information is available concerning the relative power/efficiency of the Wilcoxon and *t* tests and that the information that is available was obtained via methods whose utility has been seriously questioned. It is hoped, therefore, that the present study will both widen and strengthen this rather tentative base of knowledge.

Methodology

The present study employed large scale Monte Carlo techniques in order to study the relative power of the WSR and *t* tests in a variety of situations. In this study the power of the two tests was assessed under ten different population shapes, three sample sizes and four significance levels. Further details of these factors are provided below.

The populations chosen for study were specified in terms of their marginal rather than their difference score distributions. This was done because it was felt that researchers are more likely to have insights into the characteristics of the marginal distributions with which they are working than they are the difference score distributions that underlie their data. This is in contrast to the populations alluded to in the review of research. In those studies it was usually more convenient from a mathematical point of view to specify populations in terms of their difference score distributions.

The selection of populations for study was motivated by a desire to obtain marginal distributions that represented a wide variety of shapes. Thus distributions were chosen so as to obtain examples of symmetric, moderately skewed, highly skewed, heavy and light tailed shapes. The ten marginal distributions are listed below along with their functional forms. In addition, table 5 gives the estimated mean, variance, skew ($\beta_1^{\frac{1}{2}}$) and kurtosis (β_2) of nine of the marginal distributions as well as for the difference score distributions that are derived from

the marginals. No estimates are given for the Cauchy

Insert Table 5 About Here

---

distribution since it has infinite moments. Estimates were based on samples of size 50,000 which were generated by means of the pseudorandom number generators employed in this study. The values in Table 5 agree quite well with the known expected values of the specified marginal distributions. All populations were generated by means of the GGNML, GGUBS, GGEXN and GGCAY subroutines of the IMSL library (IMSL, 1980; see also Learmanth & Lewis, 1973). Other aspects of the study were carried out by means of a FORTRAN program written for this purpose.

normal

$$f(x)=\exp(-.5x^2)/(2\pi)^{\frac{1}{2}}, \quad -\infty<x<\infty$$

uniform

$$f(x)=1, \quad 0\leq x\leq 1$$

double exponential

$$f(x)=.5\exp(-|x|), \quad -\infty<x<\infty$$

half normal

$$f(x)=(2/\pi)^{\frac{1}{2}}\exp(-.5x^2), \quad 0\leq x<\infty$$

exponential

$$f(x)=\exp(-x), \quad 0\leq x<\infty$$

mixed normal

$$f(x)=\frac{.95}{(2\pi)^{\frac{1}{2}}}\exp(.5x^2)+\frac{.05}{(200\pi)^{\frac{1}{2}}}\exp(-(x-22)^2/200),\quad -\infty<x<\infty$$

mixed exponential

$$f(x)=.95\exp(-x)I_1(x)+.005\exp(-(x-22)/10)I_2(x)$$

where

$$I_1(x)=\begin{cases}1, & x\geq 0\\ 0, & x<0\end{cases}$$

$$I_2(x)=\begin{cases}1, & x\geq 22\\ 0, & x<22\end{cases}$$

log normal

$$f(x)=\exp(-.5(\log(x))^2)/[(2\pi)^{\frac{1}{2}}x],\quad x\geq 0$$

chi square (1 d.f.)

$$f(x)=\exp(-x/2)/(2\pi x)^{\frac{1}{2}},\quad x\geq 0$$

Cauchy

$$f(x)=1/[\pi(1+x^2)],\quad -\infty<x<\infty$$

Sample sizes investigated were 10, 25 and 50 with assessments being made at one-tailed significance levels of .05, .025, .01 and .005.

Conceptually, the algorithm used to construct the data sets for analysis is as follows. Consider a situation in which a group of n subjects has been tested at two distinct points in

time with some treatment being administered during the intervening period. The pretest score of an individual can be represented as:

$$x_{i1}=t_i+e_{i1} \qquad i = 1, 2, \ldots, n$$

where x_{i1} denotes the score of the ith subject at the first testing period, t_i the true score component of the ith subject's score and e_{i1} is the random error component of the ith subject's score at the first testing period.

An individual's posttest score can be represented as:

$$x_{i2}=c+t_i+e_{i2} \qquad i=1, 2, \ldots, n$$

Here x_{i2} denotes the score of the ith subject at the second testing period, t_i is as previously defined, and e_{i2} is the random error component of the ith subject's score at the second testing period. c is a number common to all subject's posttest scores and represents the treatment effect. The difference score upon which statistical tests for treatment effects is based is given by:

$$x_{i2}-x_{i1}=c+e_{i2}-e_{i1} \qquad i=1,2,\ldots,n.$$

In generating the difference score of an individual, e_{i1} and e_{i2} are numbers randomly sampled from a common population, that population being one of the ten mentioned above. Values of c were allowed to range from 0 (i.e. no treatment effect) to a value sufficiently large so as to produce power near one for both tests. For all sampled populations, with the exception of the Cauchy, these power functions depend only upon the value of c/σ where σ is the standard deviation of $e_{i2}-e_{i1}$.

That is, if c and σ are varied in such a way as to leave the ratio c/σ unchanged, then the power functions of the t and WSR tests will also remain unchanged. Thus comparisons of the power functions of the two statistics will apply not only to the sampled populations but also to other populations of the same type but with different variances.

In the case of the Cauchy distribution, means and standard deviations are infinite. The treatment effect, c in this case, is interpreted as the median rather than the mean of the difference score distribution. The scale of the Cauchy error distribution can be changed by multiplying each observation selected from the population by a constant k, thereby yielding power functions for the two statistics that depend only on the values of c and k through the ratio c/k. In this study k was equal to 17.

The study proceeded as follows. For each value of c, five thousand samples were generated as described above with each sample consisting of n difference scores. Wilcoxon and t statistics were computed for each of the samples and their values compared to the appropriate tabled critical values. The proportions of null hypothesis rejections obtained for the two tests at each of the four significance levels was then recorded. This procedure was carried out for each of the three sample sizes under each of the ten population shapes.

Presentation of Results

Figures 1 through 12 depict plots of the power functions of the two statistics obtained for the two population shapes. With

two exceptions, each figure shows the power functions of the two statistics obtained for samples of sizes 10, 25 and 50 under a specified distribution. The two exceptions were for the mixed normal and mixed exponential distributions. In these instances, for the sake of enhanced readability, plots for samples of size 10 and for samples of sizes 25 and 50 are presented on separate graphs. All plots are for one-tailed α=.025.

Tables 6, 7 and 8 show, for samples of sizes 10, 25 and 50 respectively, the maximum power advantage gained by each of the two statistics over the other under each of the ten populations.

Insert Tables 6, 7 and 8 About Here

For example, Figure 5 shows that for samples of size 10 drawn from an exponential distribution, the WSR test was more powerful than the t test for the smaller values of c/σ but that this situation was reversed for the higher values of c/σ. The .025 column of Table 6 shows that the maximum power advantage (i.e. the maximum vertical distance between the plots) gained by the WSR test was .021 while the maximum advantage gained by the t test was .015. These tables also give comparable values for plots obtained at α=.05, .01 and .005 although, due to space limitations, these plots are not shown. Values of .000 in these tables indicate that the specified statistic never held an advantage over the other throughout the range of the two power functions.

Results for the Normal Distribution

As expected, Figure 1 shows a clear cut albeit small, power advantage for the t test at all three sample sizes under the normal distribution. Table 6 indicates that the maximum advantage gained for n=10 at α=.025 was only .019 but was as large as .047 at α=.005. Results obtained for α=.05 are somewhat misleading since, because of the discrete nature of its sampling distribution, the WSR test was conducted at a significance level of approximately .053 rather than the desired .050. This difficulty arose only for n=10 and α=.05. The trivial advantages of .001 gained by the WSR test at the two lower significance levels are attributable to sampling errors.

Tables 7 and 8 show that the maximum advantage gained by the t for n=25 was between .014 and .017 while the comparable values for n=50 were .015 to .025. Thus it can be seen that while the t test was always the more powerful statistic when its assumptions were met, it is noteworthy that the magnitudes of its advantages were always quite small.

Results for the Uniform Distribution

Table 2 shows a general power superiority for the t test under the uniform distribution. Table 6 indicates that for samples of size 10, this advantage reached maximum values of from .031 to .052. (Results for α=.05 are being discounted because of the difficulty alluded to previously.) Table 7 shows that for samples of size 25 the maximum advantage for the t test ranged from .033 to .048. Table 8 shows that comparable values

for samples of size 50 ranged from .040 to .059.

Again it must be noted that, while the t test was the more powerful statistic under this extremely light-tailed distribution, the magnitude of its superiority was always quite modest. Interestingly, the advantage gained by the t under the uniform distribution seemed slightly higher than that attained under normal theory.

Results for the Double Exponential Distribution

Figure 3 shows that for samples of size 10 the WSR test holds a very small advantage over the t test for lower values of c/σ but that this advantage reverts to the t statistic for higher values of c/σ. Table 6 indicates that the maximum advantage for the t ranged from .008 to .030 while comparable values for the WSR ranged from .007 to .013.

Figure 6 also shows that the rather nebulous pattern for n=10 gives way to a clear advantage for the WSR test at the larger sample sizes. For samples of size 25 the maximum advantage for the WSR test ranged from .038 to .047 while comparable values for samples of size 50 ranged from .046 to .059.

Thus it can be seen that the somewhat heavy-tailed double exponential distribution tended to produce results favoring the nonparametric test although it must be pointed out that these advantages were always quite modest. It is also interesting to note the role that sample size played in determining the relative power of the two tests under this distribution. While neither test maintained a clear advantage for the smaller sample size, the WSR

dominated for the larger sample sizes.

Results for the Truncated Normal Distribution

Figure 4 shows that for samples of size 10, the WSR test may hold a very small advantage over the t for smaller values of c/σ but that the t holds a somewhat larger, although it is still small, advantage over the WSR for the larger values of c/σ. The maximum advantage for the WSR ranged from .003 to .005 while comparable values for the t ranged from .012 to .037.

Figure 4 once again indicates that the situation for the larger sample sizes is somewhat different from the situation obtained with the smaller sample size. In the former situation the WSR seems to dominate though only to a small extent. For samples of size 25 the maximum advantage of the WSR ranged from about .004 to .018. The small advantages shown for the t in this situation are probably attributable to sampling errors.

For samples of size 50, the maximum advantage for the WSR ranged from .011 to .016. The t test showed maximum advantages of from .000 to .004 which are once again probably due to sampling errors. Thus it seems that neither test held a significant advantage over the other under the truncated normal distribution although the t test faired better in the small sample situation while the same was true for the WSR in the case of the larger sample sizes.

Results for the Exponential Distribution

Figure 5 indicates that for samples of size 10 randomly

drawn from an exponential distribution, the WSR test once again holds the power advantage for small values of c/σ but that the *t* test gains the advantage for larger c/σ. Maximum advantages for the WSR ranged from.021 to .030 while those for the *t* ranged from .015 to .035.

For the larger sample sizes, the WSR test was clearly the more powerful statistic. For samples of size 25 the maximum power advantage for the WSR ranged from .087 to .100. For samples of size 50, maximum advantages for the WSR ranged from .111 to .150.

Again, the relative power of the two statistics under study was heavily influenced by sample size. Under this distribution, most observers would probably agree that the WSR test gained power advantages over the *t* test of sufficient magnitude so as to be of some practical importance in applied contexts. This latter statement is particularly credible when one considers the magnitudes of the *t* test's advantages under normal theory which appear to be the justification for the often made claim that the *t* test is more powerful than its nonparametric counterparts and should therefore be given preference in most research contexts.

Results for the Mixed Normal Distribution

Figure 6 shows the now familiar pattern for samples of size 10 whereby the WSR test dominates over lower values of c/σ while the *t* test gains the advantage for the higher values. In this situation the maximum advantage for the *t* test ranged

from .033 to .074 while comparable values for the WSR ranged from .186 to .206.

Figure 7 shows both consistent and very large power advantages for the WSR test for samples of sizes 25 and 50. For samples of 25 the *t* test never held an advantage while the WSR test held maximum advantages that ranged from .526 to .639. For samples of size 50, the maximum advantage for the WSR test ranged from .683 to .768.

For the distributions considered to this point, both tests produced obtained Type I error rates that were in reasonable agreement with nominal values. While this remained true for the WSR test under the mixed normal distribution, the *t* test produced markedly conservative Type I error rates in this situation. For n=10 the obtained probabilities for the .050, .025, .010 and .005 nominal values were respectively; .024, .011, .004 and .002. Comparable values for n=25 were .034, .009, .002 and .001 while those for n=50 were .051, .019, .004 and .001.

Under the markedly nonnormal mixed normal distribution, the nonparametric test showed truly impressive power advantages over the *t* test. While the *t* test's nonrobustness, in terms of Type I errors, to this population shape may have contributed to its distinct power disadvantage, its conservative Type I error rate does not explain this deficiency. This can be seen by noting the previously mentioned fact that for n=50 and α=.05, the *t* test produced a Type I error rate of .051 which was approximately equal to that produced by the WSR test. In spite

of these common Type I error rates, Table 8 shows that for n=50 and α=.05 the WSR test attained a maximum advantage of .683 while the *t* test never held the advantage.

Results for the Mixed Exponential Distribution

Figure 8 which represents the situation in which samples of size 10 were taken from a mixed exponential distribution, displays the same mixed pattern of power advantages found in connection with other distributions. In this case the maximum advantage for the *t* test ranged from .032 to .105 while those for the WSR test ranged from .232 to .241. Noteworthy is the fact that the value of .105 reported for the *t* test represents the largest advantage ever gained by this statistic in this study.

Figure 9, which represents the situation in which samples of sizes 25 and 50 were drawn from a mixed exponential distribution, once again reveals very large power advantages for the nonparametric test. For samples of size 25 maximum advantages for the WSR test ranged from .657 to .706 while those for n=50 ranged from .786 to .895.

In Figure 9 the power function for the *t* test at n=50 drops slightly below the function at n=25 at one point. This is attributable to sampling error and to the fact that under this distribution increasing the sample size from 25 to 50 had only minimal effects on the power of the *t* test.

Once again the *t* test showed poor agreement between obtained and nominal Type I error rates. For samples of size 10

the obtained values for the .050, .025, .010 and .005 nominal levels of significance were respectively .022, .008, .002 and .001. Comparable values for n=25 were .036, .010, .001 and .000 while those for n=50 were .049, .019, .004 and .001. As previously, the WSR test maintained reasonable agreement between obtained and nominal Type I error rates.

As with the mixed normal distribution, the nonparametric test showed extremely large power advantages over the t test under the markedly nonnormal mixed exponential distribution. Interestingly, both tests gained their greatest advantage in this study under this distribution. For the WSR test this advantage was .895 while for the t test the value was a much more modest .105.

Results for the Log Normal Distribution

As with other distributions, Figure 10 reveals a mixed pattern of advantages at n=10 that gives way to clear advantages for the WSR test at the two larger sample sizes. For samples of size 10 maximum advantages for the WSR test ranged from .058 to .072 while those for the t test ranged from .011 to .021. For samples of size 25 and 50 the WSR was always more powerful than the t test with maximum advantages at n=25 ranging from .221 to .234 and maximum advantages at n=50 ranging from .298 to .356.

For samples of size 10 the t test produced Type I error rates of .041, .017, .005 and .002 for the usual nominal levels. For samples of size 25 the comparable values were .049, .020, .006 and .002 while the values for samples of size 50 were .048,

.021, and .003. Type I error rates for the distribution-free procedure were, as expected, reasonably close to nominal values.

Though not so large as those found under the mixed normal and exponential distributions, the WSR test again showed very significant power advantages over the parametric test under the log normal distribution. The t test also produced somewhat conservative Type I error rates under this distribution.

Results for the Chi Square Distribution

Figure 11 once again reveals the now common pattern of mixed advantages at n=10 that gives way to clear advantages for the WSR test at the larger sample sizes. For samples of size 10 the maximum advantage of the t test ranged from .025 to .032 while those for the WSR test ranged from .049 to .057. For samples of size 25 maximum advantages of the WSR test ranged from .165 to .178 while comparable values for samples of size 50 ranged from .238 to .270.

As for Type I errors, the t test produced rates of .044, .018, .005 and .002 for n=10; .050, .022, .007 and .003 for n=25 and .050, .023, .007 and .004 for n=50. The WSR test again showed good agreement with nominal values.

Results for the Cauchy Distribution

As figure 12 shows, the extremely heavy-tailed Cauchy distribution produced advantages for the WSR test at all three sample sizes. For n=10 the maximum power advantage for the WSR test ranged from .109 to .113 while those for n=25 ranged from .427

to .499 and those for n=50 ranged from .659 to .701.

Type I error rates for the t test at the usual nominal levels were .025, .008, .002 and .001 for n=10. For samples of size 25 these values were .032, .010, .002 and .000 while at n=50 they were .033, .011, .003 and .001. Results for the distribution-free test were, as previously, in good agreement with nominal values.

Concluding Comments

At least insofar as the parametric paired samples t and the nonparametric WSR tests are concerned, the often repeated and too frequently unqualified claims of parametric power superiority are unjustified. In fact the results of this study, though certainly not conclusive, suggest that the asymptotic results obtained by Hodges and Lehmann (1956) carry over in some approximate sense to finite sample sizes. It will be remembered that these statisticians showed that so long as one is testing for shift, which implies a symmetric difference score population, and so long as the difference score population is continuous, its shape being otherwise unspecified, the A.R.E. of the WSR test relative to the t test may be as high as infinity but may never fall below a lower bound of .864. In the present study, in spite of the wide variety of population shapes studied, no instance was found in which the t test held more than a modest power advantage over the WSR test with the largest advantage being only .105. For the most part, in those situations in which the t test was the more powerful statistic, the magnitude of its

advantage was too small to be of much practical importance. On the other hand, the nonparametric test was not only more often the more powerful test, but in those situations where it was more powerful it was often vastly so. While the greatest advantage gained by the WSR test was .895, advantages greater than .1 were quite common.

Although the lack of citations and/or explanations make it difficult to document, it seems likely that the primary rationale used by textbook authors as the basis for their recommendation that the t test be given preference over the WSR test lies in the fact that the parametric test attains UMP status under normal theory assumptions. As this study has shown however, even when the t test's normal theory assumptions are met, the researcher has little to gain by use of the t test in place of the WSR test. The power advantages gained by the t test under the normal curve were simply too small to be of much practical importance. It seems difficult, therefore, to justify the use of the t test in situations where the shape of the sampled population is unknown on the basis of the power advantage that will be obtained if the population happens to be normal. There is simply too little to gain and too much to lose in this situation.

Although authors often present nonparametrics as "small sample techniques" and suggest that the use of such tests be largely restricted to these situations, this study has shown that there are compelling reasons for employing the WSR test in the moderate to large sample size situation. While it is true that a

researcher may choose to use the nonparametric test when sample sizes are small because of doubts concerning the robustness of the *t* test in this situation, it is also true that a researcher may choose to use the nonparametric test in the larger sample size situation in order to realize the power advantages that often accrue to this statistic in this circumstance.

Three final points are in order. First, the objection might be raised that rarely are data gathered in the course of social and behavioral science inquiries sufficiently nonnormal to warrant concern. In response we would modify slightly a statement by Blair (1981) and state the following, "Perhaps the most effective means of dealing with such a notion on the part of...[a researcher in the social and/or behavioral sciences]...is to suggest that he/she routinely construct relative frequency histograms of the data that they submit to statistical analysis. This time-honored but often neglected practice usually paints pictures of distributions that are unimagined by researchers who think of data in terms of the normal curve." Indeed, it was concern brought about by the appearance of nonnormal data that motivated the early robustness studies (Boneau,1960). Readers interested in examples and discussions of nonnormal data in the social and behavioral sciences should see Blair (1981), Bradley (1977) and Walberg, Strykowski, Rovai and Hung (1984) as well as references in the two latter articles.

Second, some authors seem to advocate the exclusive use of parametric procedures even in situations where it is known that the assumption of population normality has been rather seriously violated (Glass, Peckham and Sanders, 1972). The reader should clearly understand that the results of this study do not support a mirror-image argument to that put forth by such writers. That is, there is no basis for an argument that the WSR test should be used to the exclusion of the t test in data analysis. Such an argument is short sighted and unfounded regardless of whether it is the parametric or nonparametric test that is being advocated. It must be remembered that the present study was designed to assess the effects of population shape on the power of the two statistics when testing for shift. Situations not investigated in this study are certain to influence the relative power of the two tests.

For example, since the t test is necessarily the test of choice when populations are normally distributed only if one is testing for shift and since few other treatment effect models have been proposed in the social and behavioral sciences, this seemed a reasonable starting point for a study such as this. However, as new treatment effect models are proposed and validated in various disciplines, it is likely that these models will significantly effect the power of these as well as other tests. We may well discover that none of the currently popular tests are as effective as we now seem to believe. In any case, this study sheds light on only one aspect of the problem.

One must also remember that the WSR test too has its assumptions . When large proportions of tied observations are present or when one is concerned about the WSR's assumption of difference score population symmetry, the t test may be preferable. In addition, some researchers may feel uncomfortable with the nature of the hypothesis being tested by the nonparametric test. While the present authors believe that too much has traditionally been made of this latter point, it nevertheless must be taken into account when choosing a test of significance.

Third, while this study dealt solely with the paired samples t test and Wilcoxon's signed-ranks test, it should be viewed as part of a small but growing body of evidence that is seriously challenging traditional views of nonparametric statistics. For example, the results reported here agree quite well with those obtained by Blair and Higgins (1980) who studied the relative power of the independent means t test and Wilcoxon's rank sum test. In addition W.J. Conover and Ronald L. Iman have shown that statistical procedures such as ANCOVA, discriminant analysis, factorial ANOVA and others when performed on ranks very often out perform their parametric counterparts that are based on original observations. For a discussion of these procedures and additional references, see Conover and Iman (1981).

In conclusion, we would strongly recommend that authors of statistical texts, as well as others, refrain from discouraging the use of nonparametric tests on the basis of unqualified and unsupported assertions of nonparametric power inferiority. When

information is available concerning the relative power of specific tests being discussed by a textbook writer, this information should be passed along to the reader. When little or no information is available concerning the relative power of specific tests, the reader should be apprised of this fact also. Decisions concerning the relative usefulness of various statistical techniques in a particular situation must be made on the basis of a variety of considerations. It is hoped that the results of this study will strengthen the knowledge base from which researchers act in order to select the most appropriate technique available for a particular problem.

References

Arnold, H. (1965). Small sample power of the one sample Wilcoxon test for non-normal shift alternatives. Annals of Mathematical Statistics, 36, 1767-1778.

Ary, D., & Jacobs, L. (1976). Introduction to statistics. New York: Holt, Rinehart and Winston.

Bahadur, P.R. (1960). Simultaneous comparisons of the optimum and sign tests of a normal mean. In I. Olkin, S.G. Ghurye, W. Hoeffding, W.G. Madow & H.B. Mann (Eds.), Contributions to probability and statistics; Essays in honor of Harold Hotelling (pp. 79-88). Stanford, CA: Stanford University Press.

Best, J. (1981). Research in education (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Blair, R.C. (1981). A reaction to "Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance". Review of Educational Research, 51, 499-507.

Blair, R.C., & Higgins, J.J. (1980). A comparison of the power of Wilcoxon's rank-sum statistic to that of Student's t statistic under various non-normal distributions. Journal of Educational Statistics, 5, 309-335.

Boneau, C.A. (1960). The effects of violations of assumptions underlying the t test. Psychological Bulletin, 57, 49-64.

Bradley, J.V. (1968). Distribution-free statistical tests. Englewood Cliffs, NJ: Prentice-Hall.

Bradley, J.V. (1977). A common situation conducive to bizarre distribution shapes. The American Statistician, 31, 147-150.

Chase, C. (1976). Elementary statistical procedures (2nd ed.). New York: McGraw-Hill.

Conover, W.J., & Iman, R.L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. The American Statistician, 35, 124-128.

Cornell, F. (1956). The essentials of educational statistics. New York: Wiley.

Couch, J. (1982). Fundamentals of statistics for the behavioral sciences. New York: St. Martin's Press.

Downie, N., & Heath, R. (1970). Basic statistical methods (3rd ed.). New York: Harper & Row.

Downie, N., & Starry, A. (1977). Descriptive and inferential statistics. New York: Harper & Row.

Elashoff, J., & Elashoff, R. (1978). Effects of errors in statistical assumptions. In W. Kruskal & J. Tanur (Eds.), International encyclopedia of statistics (pp. 229-230). New York: The Free Press.

Elifson, K., Runyon, R., & Haber, A. (1982). Fundamentals of social statistics. Reading, MA: Addison-Wesley.

Engelhart, M. (1972). Methods of educational research. Chicago: Rand McNally.

Garrett, H. (1966). Statistics in psychology and education (6th ed.). New York: David McKay.

Gay, L. (1976). Educational research. Columbus, OH: Charles E. Merrill.

Glass, G., Peckham, P., & Sanders, J. (1972). Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. Review of Educational Research, 42, 237-288.

Guilford, J., & Fruchter, B. (1978). Fundamental statistics in psychology and education (6th ed.). New York: McGraw-Hill.

Hays, W. (1963). Statistics. New York: Holt, Rinehart and Winston.

Hettmansperger, T., & McKean, J. (1978). Statistical inference based on ranks. Psychometrika, 43, 69-79.

Hinkle, D., Wiersman, W., & Jurs, S. (1982). Basic behavioral statistics. Boston: Houghton Mifflin.

Hodges, J., & Lehmann, E. (1956). The efficiency of some nonparametric competitors of the t-test. Annals of Mathematical Statistics, 27, 324-335.

Hopkins, K., & Glass, G. (1978). Basic statistics for the behavioral sciences. Englewood Cliffs, NJ: Prentice-Hall.

International mathematical and statistical libraries. (1980). Houston, TX: International Mathematical and Statistical Libraries.

Johnson, M. (1977). A review of research methods in education. Chicago: Rand McNally.

Kerlinger, F.N. (1964). Foundations of behavioral research. New York: Holt, Rinehart and Winston.

Kirk, R. (1968). Experimental design: Procedures for the behavioral sciences. Belmont, CA: Brooks/Cole.

Klotz, J. (1963). Small sample power and efficiency for the one sample Wilcoxon and normal scores tests. Annals of Mathematical Statistics, 34, 624-632.

Klotz, J. (1965). Alternative efficiencies for signed rank tests. Annals of Mathematical Statistics, 36, 1759-1766.

Klugh, H. (1970). Statistics: The essentials for research. New York: John Wiley & Sons.

Kurtz, A., & Mayo, S. (1979). Statistical methods in education and psychology. New York: Spring-Verlag New York.

Kurtz, N. (1983). Introduction to social statistics. New York: McGraw-Hill.

Learmonth, G., & Lewis, P. (1973) Statistical tests of some widely used and recently proposed uniform random number generators (NPS55LW73111A). Monterey, CA: Naval Postgraduate School.

Lehmann, E.L. (1975). Nonparametrics, statistical methods based on ranks. San Francisco: Holden-Day.

Mendenhall, W. (1971). Introduction to probability and statistics (3rd ed.). CA: Duxbury Press.

Minium, E. (1970). Statistical reasoning in psychology and education. New York: John Wiley & Sons.

Nunnally, J. (1975). Introduction to statistics for psychology and education. New York: McGraw-Hill.

Pagano, R. (1981). Understanding statistics. St. Paul, MN: West.

Roscoe, J. (1975). Fundamental research statistics for the behavioral sciences (2nd. ed.). New York: Holt, Rinehart and Winston.

Sharp, V. (1979). Statistics for the social sciences. Boston: Little, Brown and Company.

Siddiqui, M., & Raghunandanan, K. (1967). Asymptotically robust estimators of location. Journal of the American Statistical Association, 62, 950-953.

Spence, J., Underwood, B., Duncan, C., & Cotton, J. (1968). Elementary statistics. New York: Appleton-Century-Crofts.

Thompson, R., Govindarajulu, Z., & Doksum, K. (1967). Distribution and power of the absolute normal scores test. Journal of the American Statistical Association, 62, 966-975.

Walberg, H. J., Strykowski, B. F., Rovai, E., & Hung, S.S. (1984). Exceptional performance. Review of Educational Research, 54, 87-112.

Table 1

A.R.E.s of the WSR Test Relative to the t Test Under a Variety of Distribution Shapes

Distributions of difference scores	A.R.E.
uniform (or rectangular)	1.00
normal	.96
logistic	1.10
mixed normal	
k=3, T=.01	1.01
.05	1.20
.10	1.37
.20	1.57
Student's *t*	
d.f.=1	∞
3	1.90
5	1.24
double exponential	1.50

Table 2

Bahadur Relative Efficiencies of the WSR Test Relative to the t Test Under a Normal Distribution

mean	.125	.50	.75	1.00	2.00	3.00
efficiency	.957	.974	.981	.969	.784	.598

Table 3

Small Sample Efficiencies of the WSR Test Relative to the t Test Under a Normal Distribution

	mean		
n	.25	1.0	1.5
5	.986	.981	.979
α=.0625			
8	.980	.973	.971
α=.0546			
10	.968	.965	.964
α=.0257			

Table 4

Power of the Sign, WSR and t Tests When Samples Are Taken From a Cauchy Distribution

median	.25			1.0			2.0		
	s	WSR	t	s	WSR	t	s	WSR	t
n=5	.34	.33	.25	.76	.76	.69	.88	.88	.83
α=.0625									
n=8	.43	.35	.26	.91	.76	.71	.98	.86	.85
α=.0547									
n=10	.50	.42	.26	.96	.90	.72	.99	.87	.84
α=.0527									

Table 5

Estimates of Mean, Variance, Skewness and Kurtosis of Nine Marginal and Difference Score Distributions

Marginal distribution	Sampled distribution	Mean	Variance	Skewness	Kurtosis
normal	marginal	-.009	.988	.011	3.05
	diff. score	-.009	1.983	.005	3.02
uniform	marginal	.501	.083	.000	1.887
	diff. score	.001	.167	.000	2.402
double exponential	marginal	-.001	1.979	-.003	5.941
	diff. score	-.001	3.985	.001	4.502
truncated normal	marginal	.795	.363	.983	3.787
	diff. score	-.003	.722	-.007	3.453
exponential	marginal	1.000	1.002	1.989	8.842
	diff. score	.000	1.991	-.018	5.981
mixed normal	marginal	1.116	29.752	4.907	28.713
	diff. score	.032	58.391	.113	15.844
mixed exponential	marginal	2.571	51.748	4.802	29.086
	diff. score	.030	103.776	-.052	16.361
log normal	marginal	1.641	4.605	6.644	128.607
	diff. score	-.003	9.103	.371	50.881
chi square	marginal	.985	1.915	2.726	13.768
	diff. score	-.022	1.914	-.062	8.480

Table 6

Maximum Power Advantages Gained by the t and Wilcoxon Statistics Under Ten Population Shapes, n=10

		One-tailed significance levels			
Population	Statistic	.050	.025	.010	.005
normal	t	.006	.019	.028	.047
	WSR	.006	.000	.001	.001
uniform	t	.020	.031	.033	.052
	WSR	.007	.000	.000	.000
double exponential	t	.000	.008	.020	.030
	WSR	.024	.009	.007	.013
truncated normal	t	.003	.012	.029	.037
	WSR	.016	.005	.004	.003
exponential	t	.000	.015	.029	.035
	WSR	.045	.021	.024	.030
mixed normal	t	.000	.074	.042	.033
	WSR	.309	.186	.206	.186
mixed exponential	t	.000	.105	.071	.032
	WSR	.383	.232	.239	.241
log normal	t	.000	.021	.020	.011
	WSR	.104	.058	.063	.072
chi square	t	.000	.026	.032	.025
	WSR	.086	.049	.050	.057
Cauchy	t	.000	.000	.000	.000
	WSR	.221	.110	.109	.113

Table 7

Maximum Power Advantages Gained by the t and Wilcoxon Statistics Under Ten Population Shapes, n=25

		One-tailed significance levels			
Population	Statistic	.050	.025	.010	.005
normal	t	.015	.014	.014	.017
	WSR	.000	.004	.000	.000
uniform	t	.037	.033	.038	.048
	WSR	.000	.001	.001	.000
double exponential	t	.000	.000	.000	.000
	WSR	.038	.047	.043	.041
truncated normal	t	.001	.001	.000	.003
	WSR	.014	.018	.011	.004
exponential	t	.000	.000	.000	.000
	WSR	.087	.100	.099	.093
mixed normal	t	.000	.000	.000	.000
	WSR	.526	.580	.637	.639
mixed exponential	t	.000	.000	.000	.000
	WSR	.657	.681	.705	.706
log normal	t	.000	.000	.000	.000
	WSR	.221	.233	.234	.227
chi square	t	.000	.000	.000	.000
	WSR	.178	.176	.172	.165
Cauchy	t	.000	.000	.000	.000
	WSR	.477	.499	.483	.427

Table 8

Maximum Power Advantages Gained by the t and Wilcoxon Statistics Under Ten Population Shapes, n=50

		One-tailed significance levels			
Population	Statistic	.050	.025	.010	.005
normal	t	.025	.015	.024	.021
	WSR	.000	.001	.002	.002
uniform	t	.042	.040	.053	.059
	WSR	.000	.000	.000	.000
double exponential	t	.000	.000	.000	.000
	WSR	.046	.056	.059	.055
truncated normal	t	.004	.001	.002	.000
	WSR	.016	.016	.011	.016
exponential	t	.000	.000	.000	.000
	WSR	.115	.129	.111	.150
mixed normal	t	.000	.000	.000	.000
	WSR	.683	.747	.768	.760
mixed exponential	t	.000	.000	.000	.000
	WSR	.786	.822	.817	.895
log normal	t	.000	.000	.000	.000
	WSR	.298	.324	.351	.356
chi square	t	.000	.000	.000	.000
	WSR	.238	.262	.268	.270
Cauchy	t	.000	.000	.000	.000
	WSR	.659	.701	.688	.670

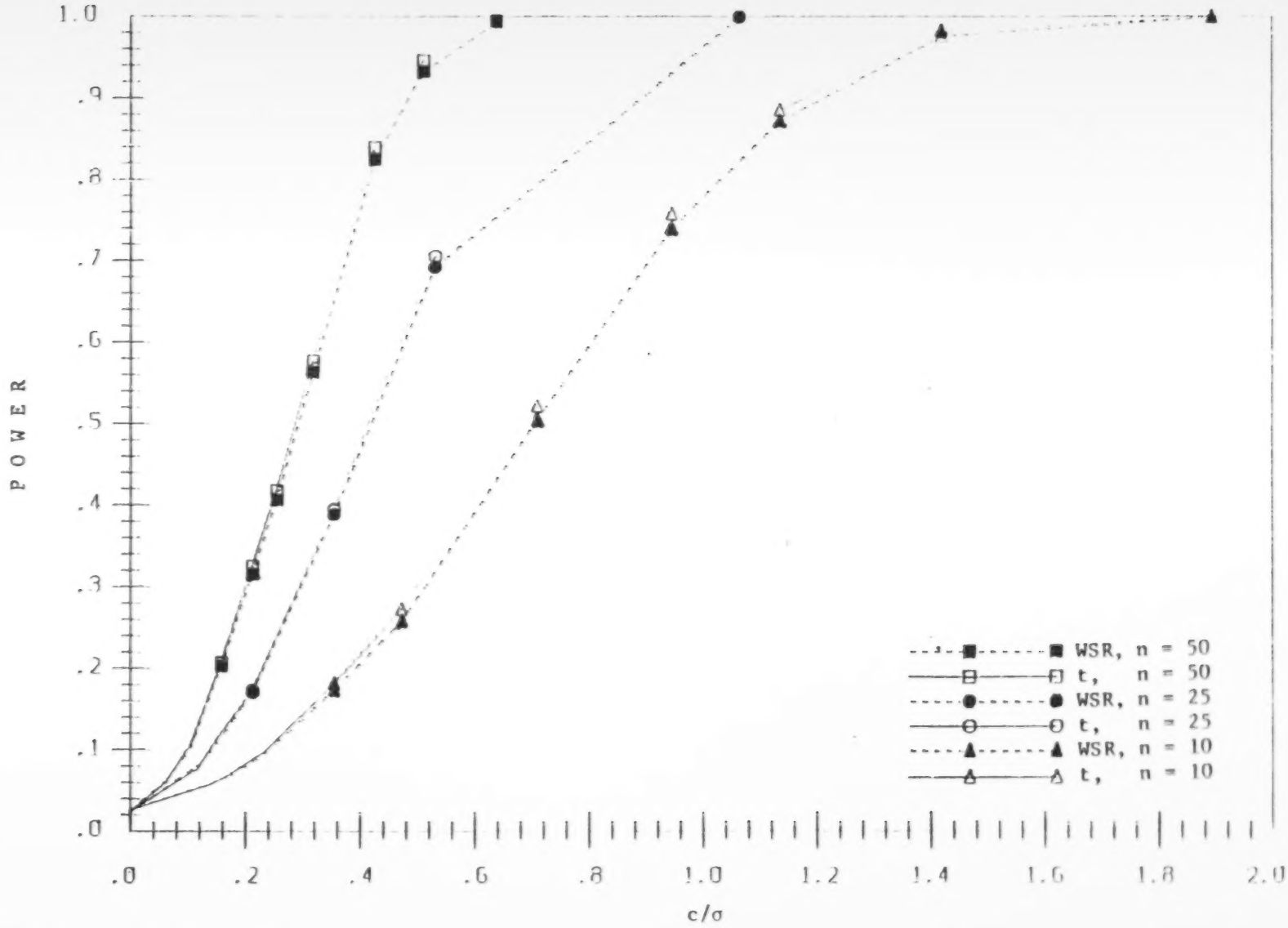

Figure 1. Power of the t and Wilcoxon signed-rank tests for samples of sizes 10, 25 and 50 randomly drawn from normal distributions.

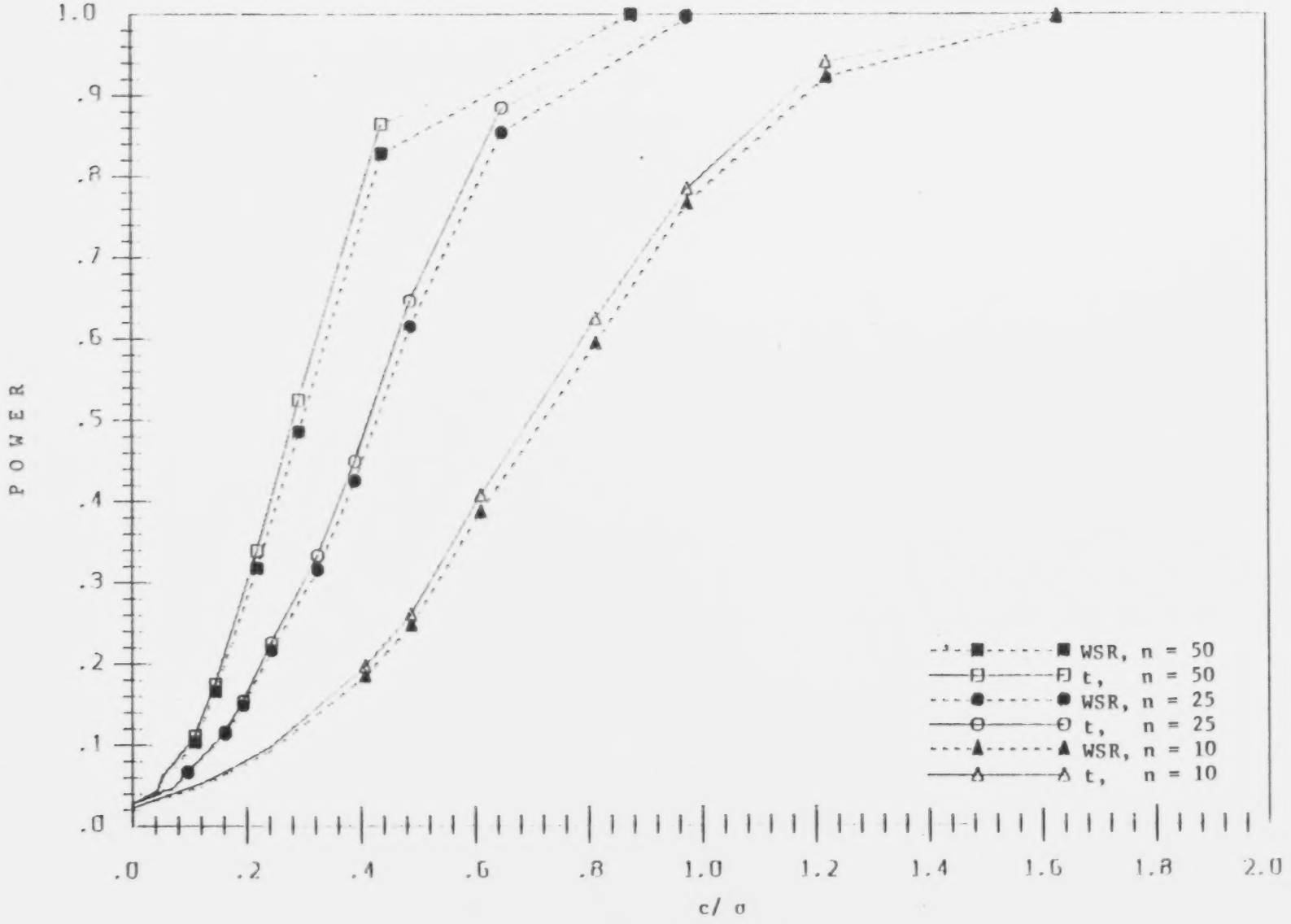

Figure 2. Power of the t and Wilcoxon signed-rank tests for samples of sizes 10, 25 and 50 randomly drawn form uniform distributions.

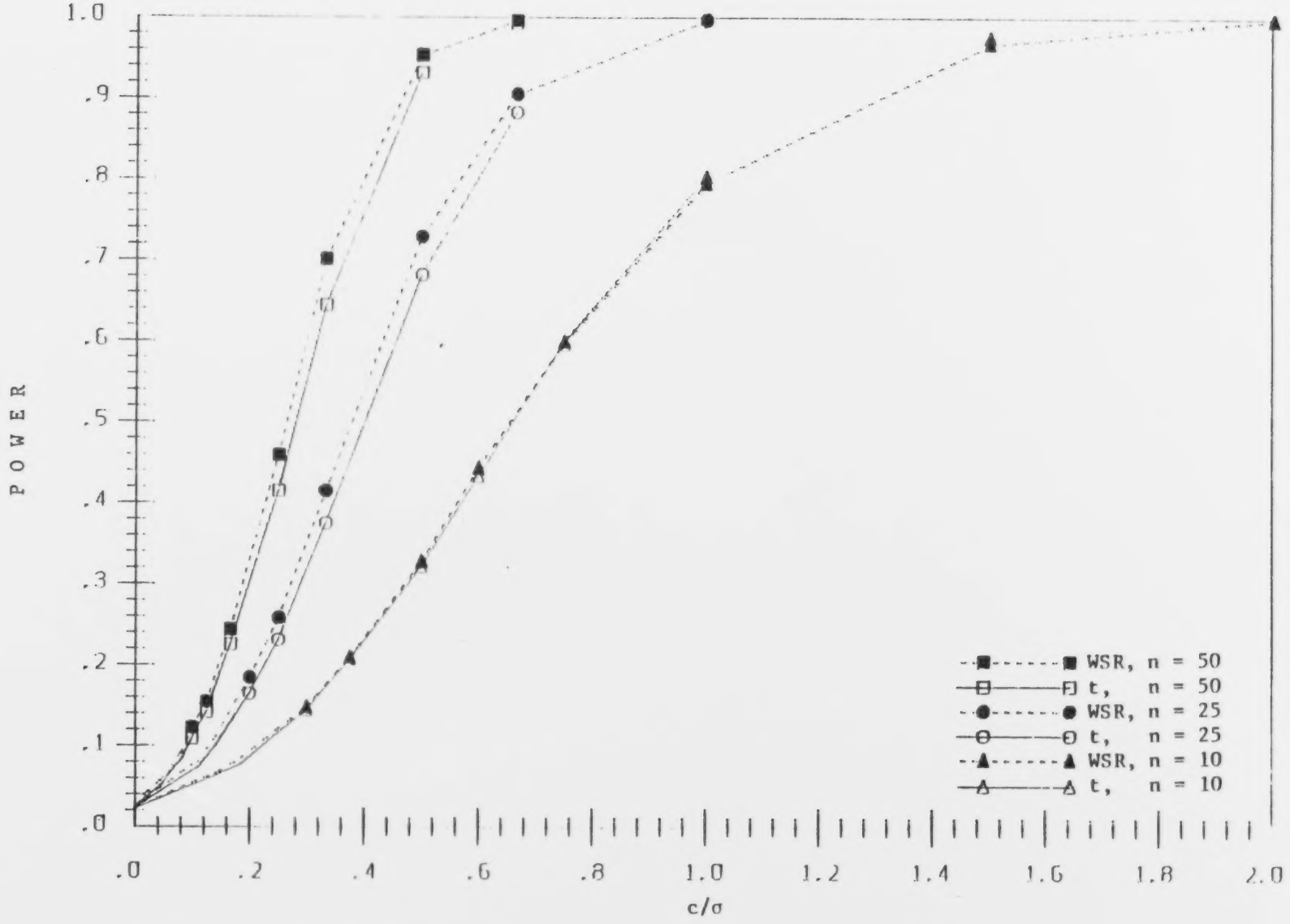

Figure 3. Power of the $\underline{t}$ and Wilcoxon signed-rank tests for samples of sizes 10, 25, and 50 randomly drawn from double exponential distributions.

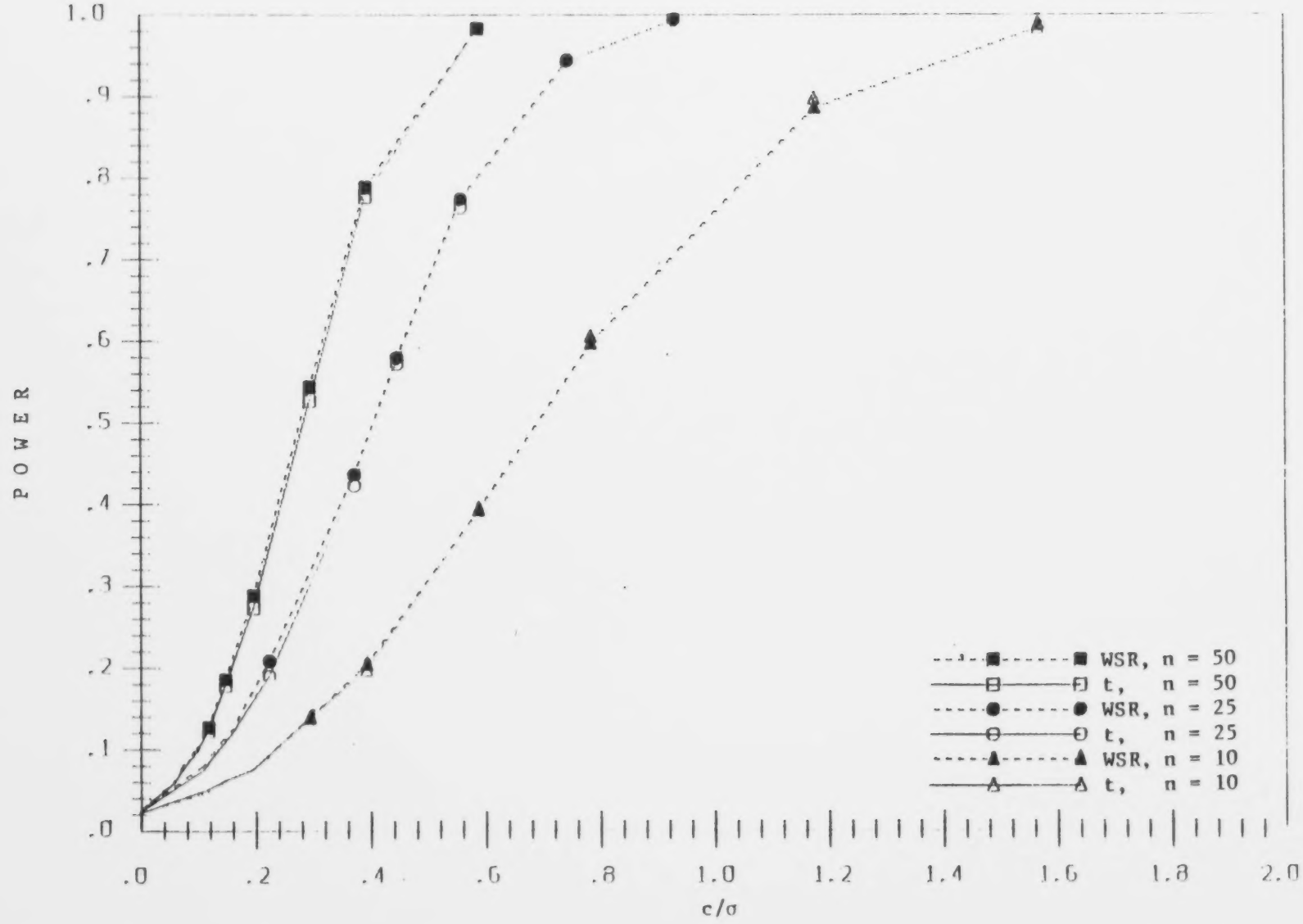

Figure 4. Power of the t and Wilcoxon signed-rank tests for samples of sizes 10, 25 and 50 randomly drawn from truncated normal distributions.

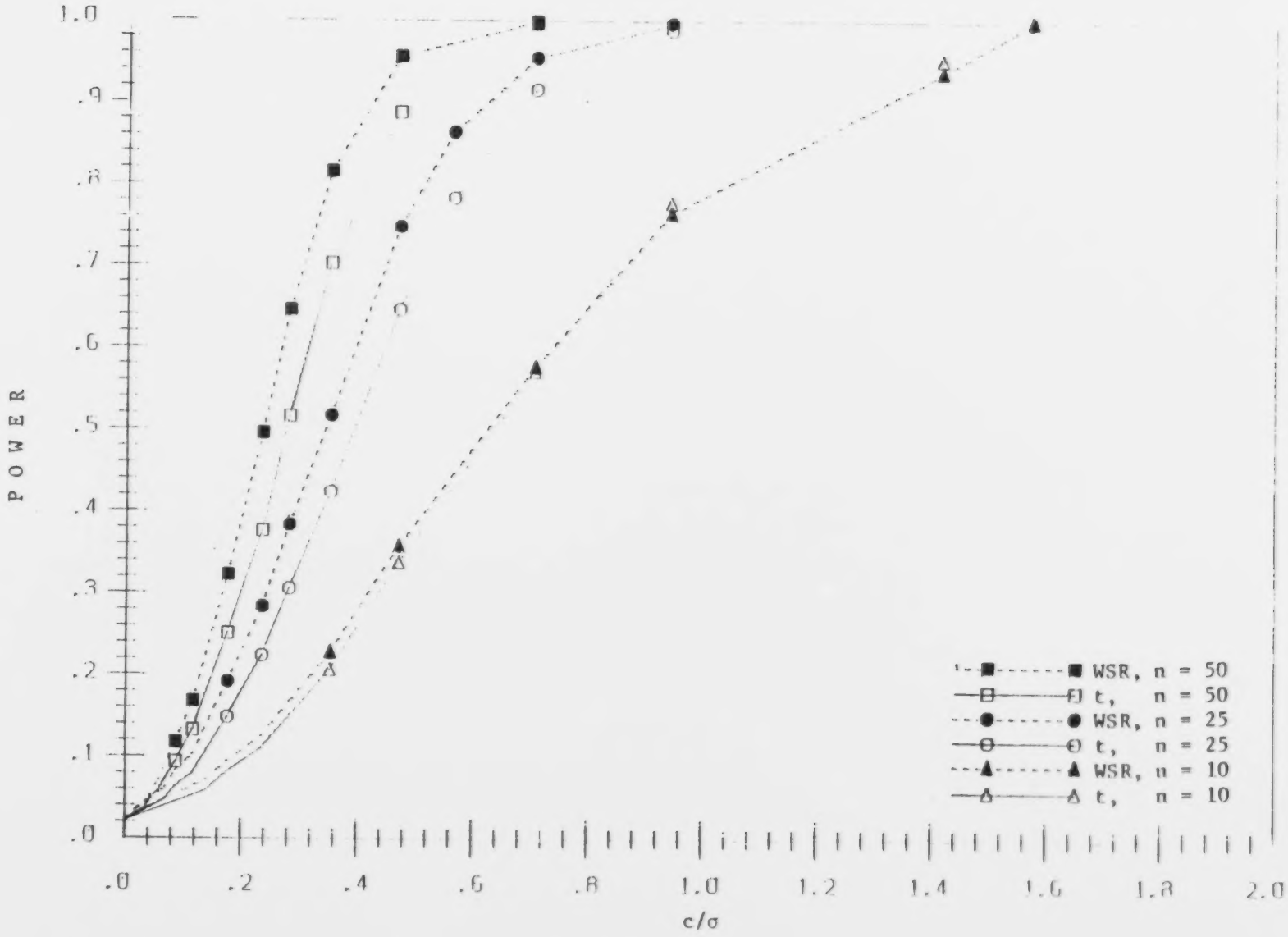

Figure 5. Power of the t and Wilcoxon signed-rank tests for samples of sizes 10, 25 and 50 randomly drawn from exponential distributions.

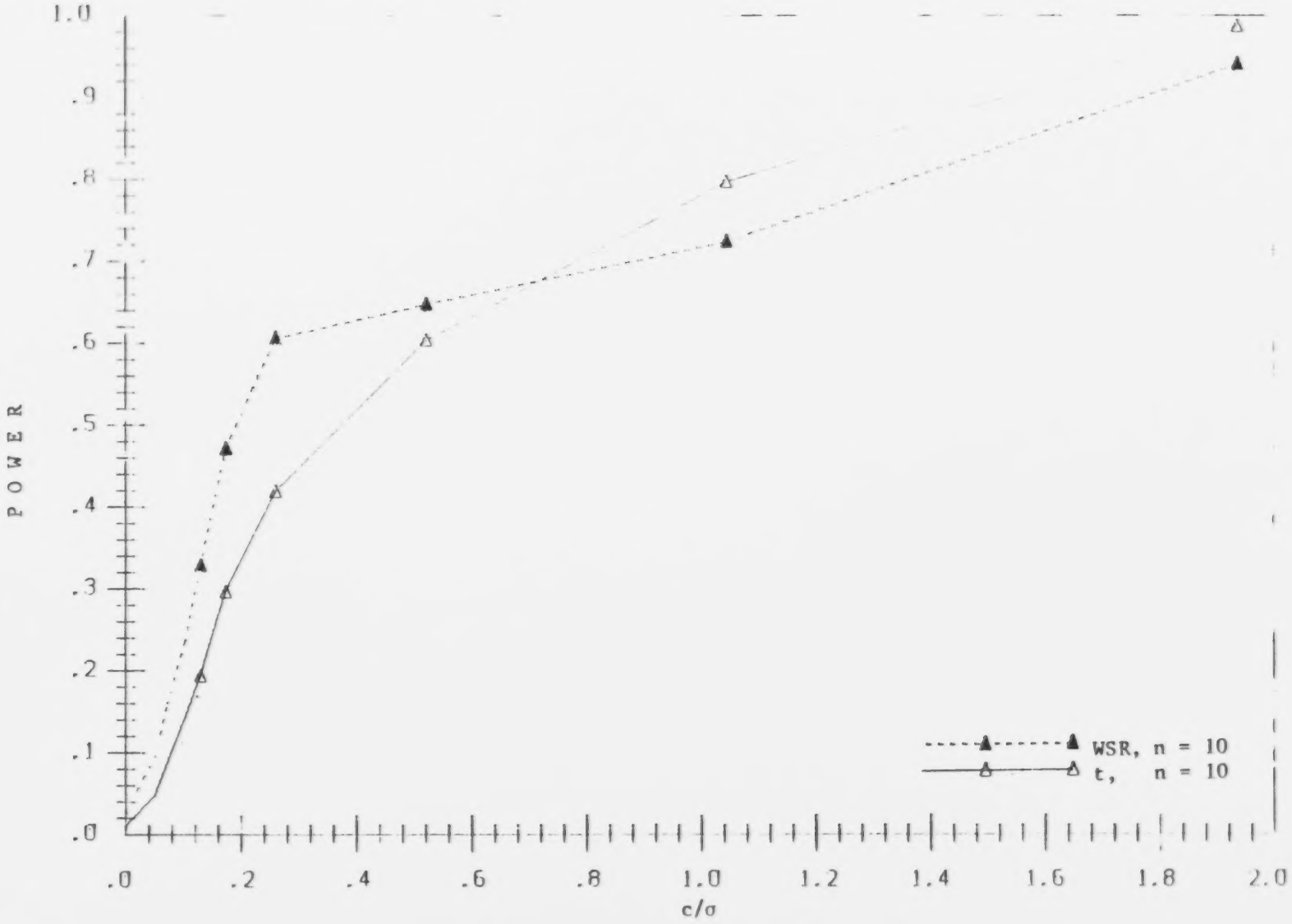

Figure 6. Power of the t and Wilcoxon signed-rank tests for samples of size 10 randomly drawn from mixed normal distributions.

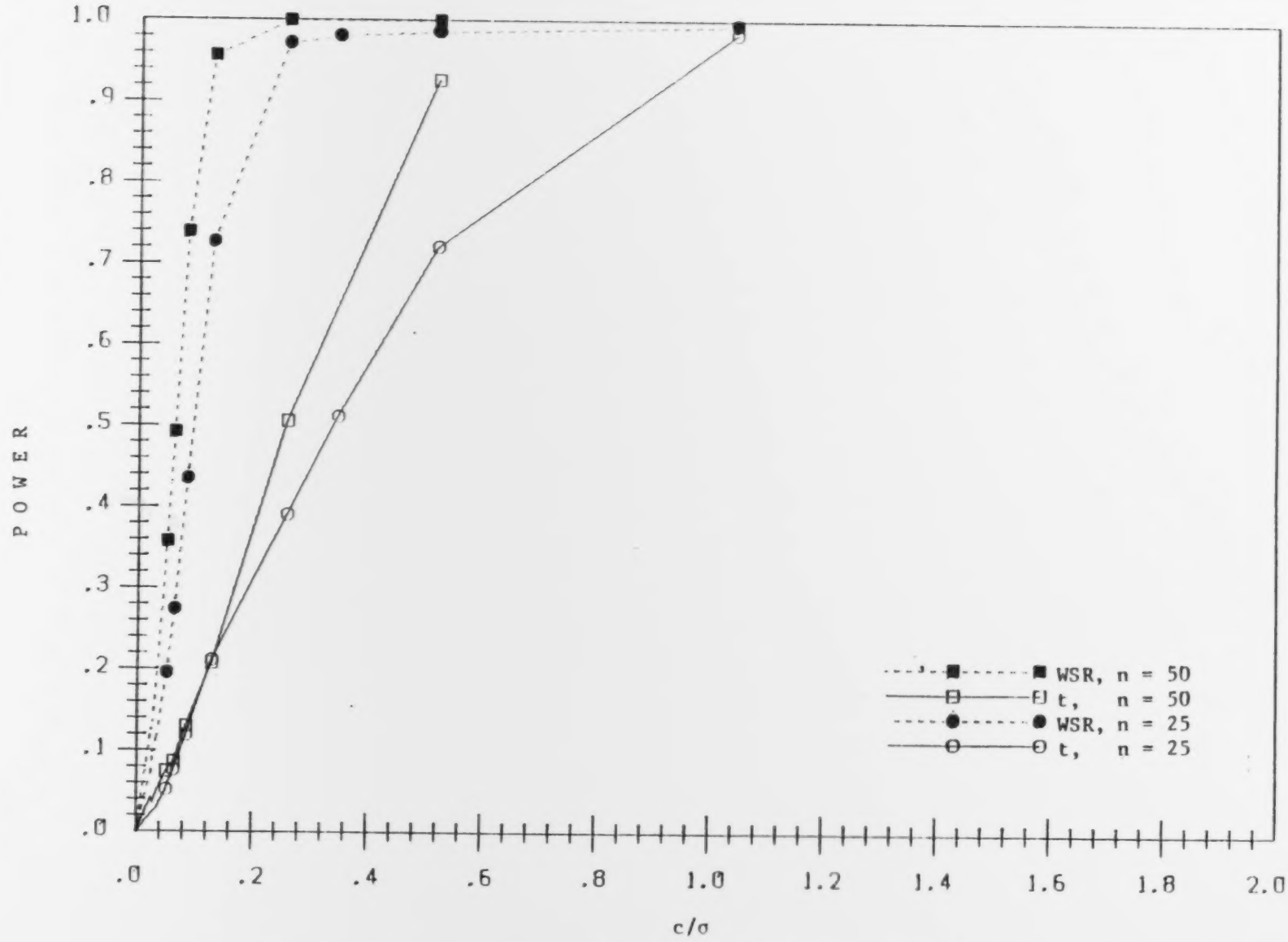

Figure 7. Power of the t and Wilcoxon signed-rank tests for samples of sizes 25 and 50 randomly drawn from mixed normal distributions.

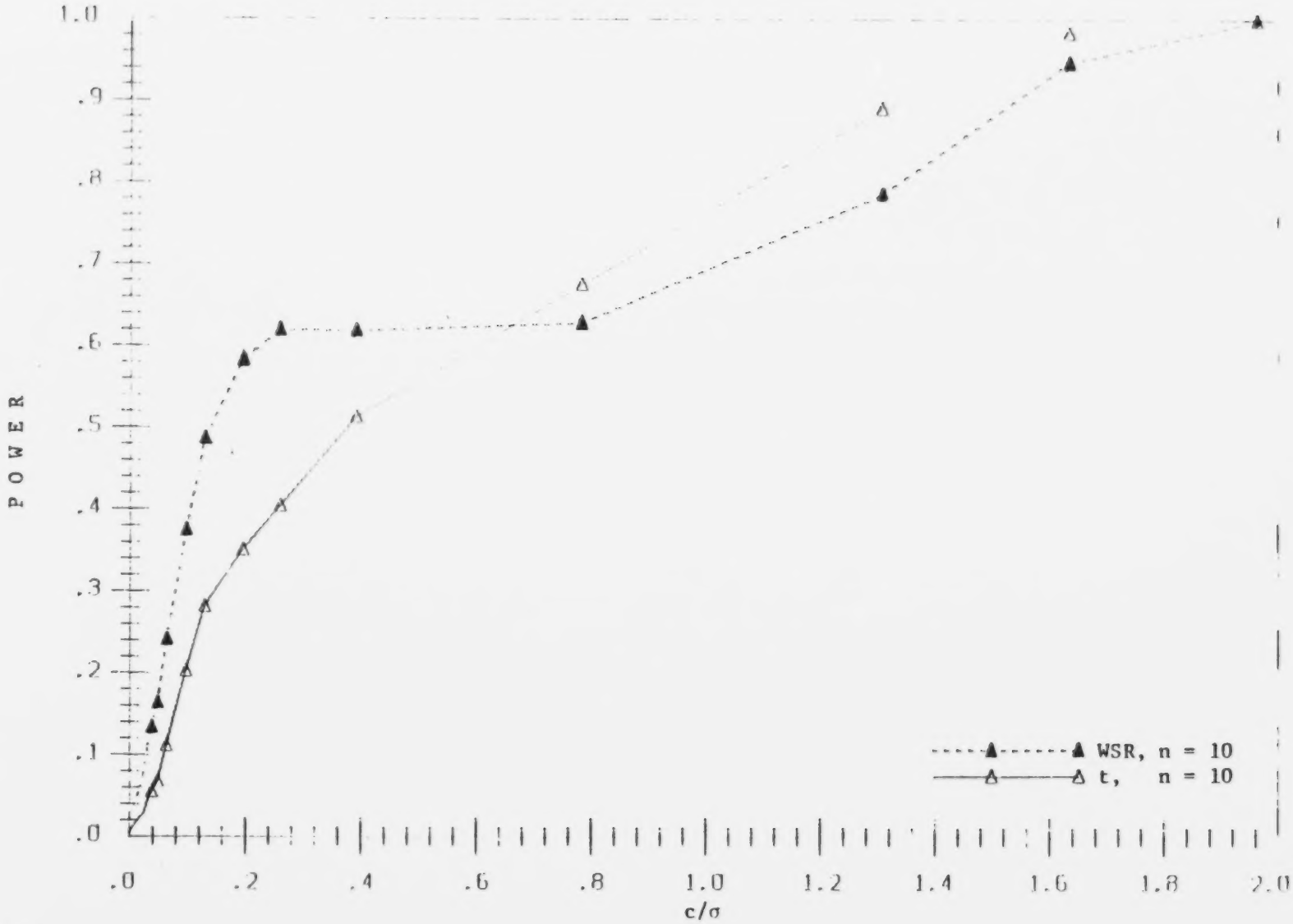

Figure 8. Power of the t and Wilcoxon signed-rank tests for samples of size 10 randomly drawn from mixed exponential distributions.

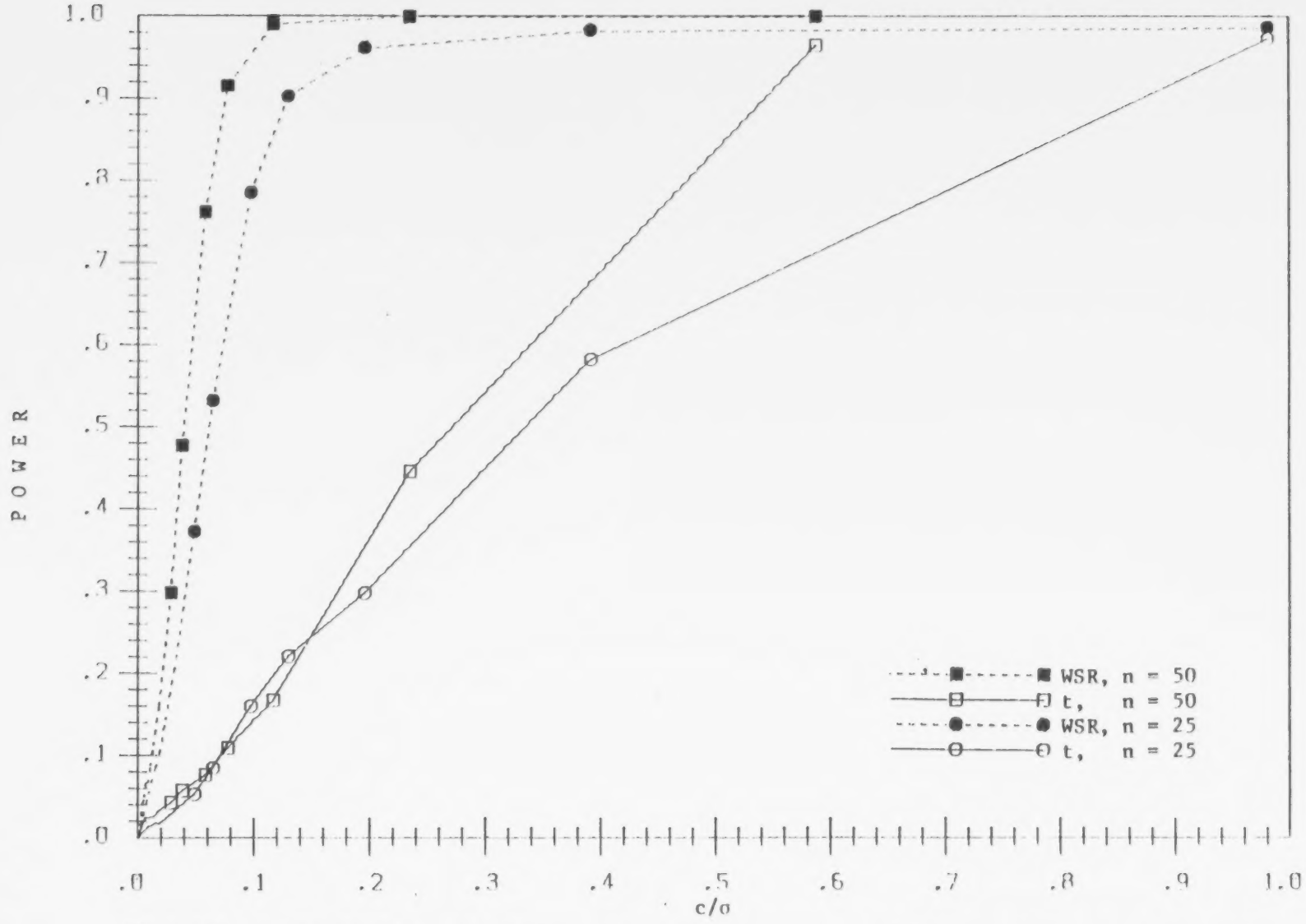

Figure 9. Power of the t and Wilcoxon signed-rank tests for samples of sizes 25 and 50 randomly drawn from mixed exponential distributions.

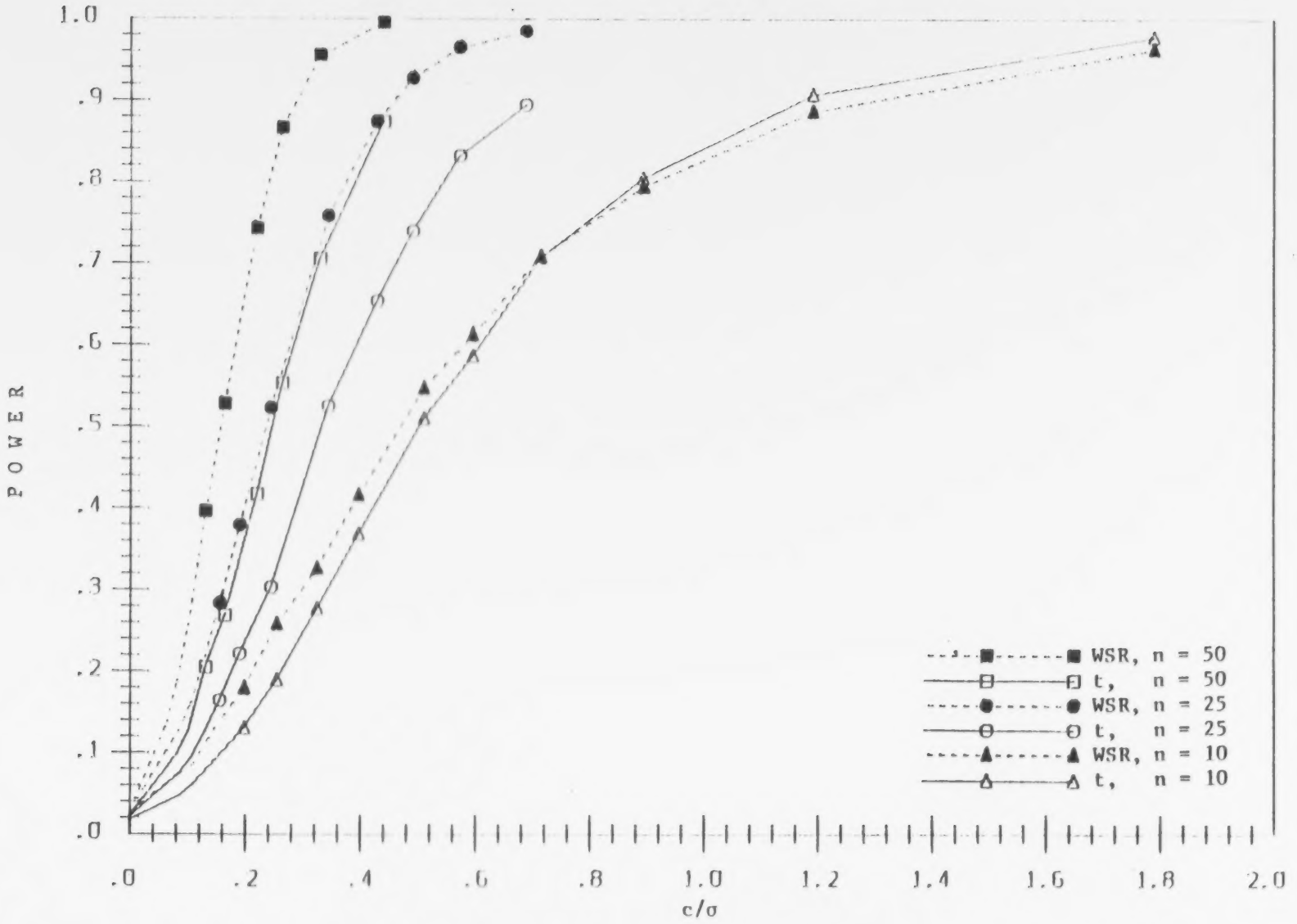

Figure 10. Power of the t and Wilcoxon signed-rank tests for samples of sizes 10, 25 and 50 randomly drawn from log normal distributions.

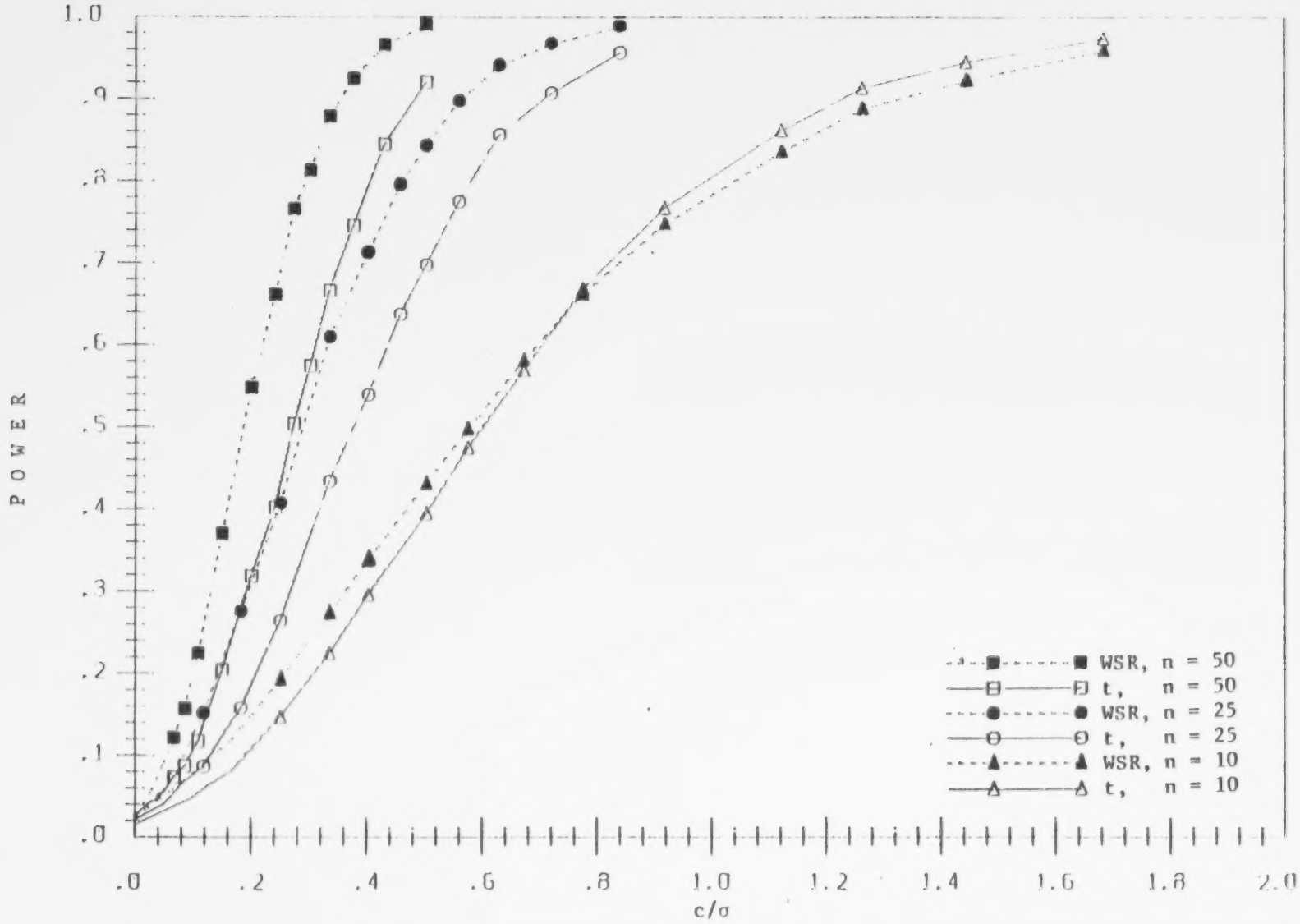

Figure 11. Power of the t and Wilcoxon signed-rank tests for samples of sizes 10, 25 and 50 randomly drawn from chi square distributions (d.f. = 1).

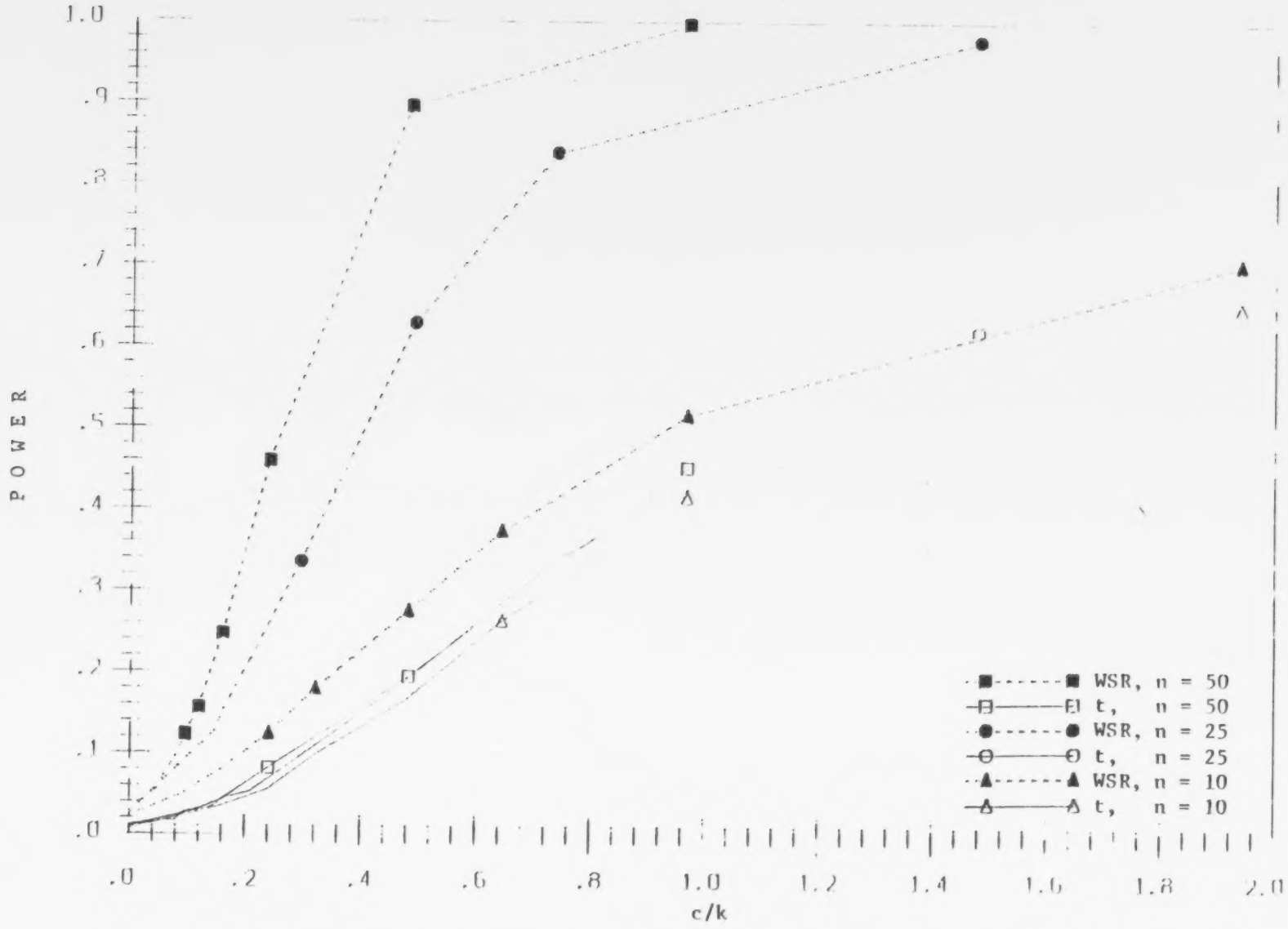

Figure 12. Power of the t and Wilcoxon signed-rank tests for samples of sizes 10, 25 and 50 randomly drawn from Cauchy distributions.